Final Exam – CS 526 – CE4

There are 4 general problems and 4 multiple choice questions. Good luck!

Name: ________________________
Section: ____________
Sciper No.: ____________

<table>
<thead>
<tr>
<th>Problem</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem 1</td>
<td>/ 20</td>
</tr>
<tr>
<td>Problem 2</td>
<td>/ 20</td>
</tr>
<tr>
<td>Problem 3</td>
<td>/ 20</td>
</tr>
<tr>
<td>Problem 4</td>
<td>/ 20</td>
</tr>
<tr>
<td>Problem 5: MCQ</td>
<td>/ 20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>/100</td>
</tr>
</tbody>
</table>
Problem 1. VC Dimension (20 pts)

In this problem we consider hypothesis functions from \( \mathbb{R}^2 \) to \( \{0, 1\} \). We have seen in the homework that \( \text{VCdim}(\mathcal{H}_{\text{rec}}) = 4 \), where \( \mathcal{H}_{\text{rec}} \) is the class of all rectangles in \( \mathbb{R}^2 \). Let us see some other examples.

1. (10 pts) (Circles) Let \( \mathcal{H}_1 = \{h_C(x)\} \) with \( h_C(x) = \mathbb{I}(x \text{ is inside the circle } C) \), where a circle \( C \) is determined by a center and a radius.

   (a) (3 pts) What is \( \text{VCdim}(\mathcal{H}_1) \)? Call your answer \( d_1 \).

   (b) (3 pts) Show that \( \text{VCdim}(\mathcal{H}_1) \geq d_1 \).

      (Hint: You can propose an instance of \( d_1 \) points and for each labelling draw the valid circle.)

   (c) (4 pts) Show that \( \text{VCdim}(\mathcal{H}_1) \leq d_1 \).

      Hint: You should consider two cases:
      - one of the points \( x \) is in the convex hull of the other points; OR
      - none of the points is in the convex hull of the other points.

      A formal proof might be difficult. It will suffice if you give us a “convincing” argument.

2. (10 pts) (Unbiased neurons) Let \( \mathcal{H}_2 = \{h_{\alpha_1, \alpha_2}(x) : \alpha_1, \alpha_2 \in \mathbb{R}\} \) with

   \[
h_{\alpha_1, \alpha_2}(x) = \mathbb{I}( \tanh(\alpha_2 x_2 + \alpha_1 x_1) > 0 ).
   \]

   (a) (3 pts) What is \( \text{VCdim}(\mathcal{H}_2) \)? Call your answer \( d_2 \).

   (b) (3 pts) Show that \( \text{VCdim}(\mathcal{H}_2) \geq d_2 \).

   (c) (4 pts) Show that \( \text{VCdim}(\mathcal{H}_2) \leq d_2 \).
Problem 2. GD and SGD (20 pts)

1. (15 pts) Consider the Least Squares optimization problem:
\[ x^* = \arg\min_{x \in \mathbb{R}^n} f(x), \]
where \( f(x) = \frac{1}{2}||Ax - b||_2^2, b \in \mathbb{R}^m \). We assume that \( A \) is a full column rank matrix in \( \mathbb{R}^{m \times n}, n \leq m \), and that there exists a solution to the linear system \( Ax = b \). Let \( \sigma_{\max} \) and \( \sigma_{\min} \) be the largest and the smallest singular values of \( A \) and consider the gradient descent method
\[ x^{t+1} = x^t - \alpha \nabla f(x^t) \]
with a fixed step size \( \alpha = 1/\sigma_{\max}(A)^2 \).

(a) (5 pts) Show that \( \sigma_{\max}(I - \alpha A^T A) = 1 - \alpha \sigma_{\min}(A)^2 = 1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2} \).

(b) (5 pts) Calculate the gradient \( \nabla f(x) \) and rewrite the GD using this gradient.

(c) (5 pts) Show that the procedure converges as
\[ ||x^{t+1} - x^*||_2 \leq (1 - \frac{\sigma_{\min}(A)^2}{\sigma_{\max}(A)^2}) ||x^t - x^*||_2. \]

2. (5 pts) Let us now consider the SGD. In this case one can show a convergence of the form
\[ \mathbb{E}[||x^{t+1} - x^*||_2^2] \leq (1 - \frac{\sigma_{\min}(A)^2}{||A||_F^2}) \mathbb{E}[||x^t - x^*||_2^2] \]
where \( ||A||_F \) is the Frobenius norm. How does this compare to GD? Which is better?
Problem 3. Probabilistic graphical models (20 pts)

Let $X_t$, $t = 0, 1, 2$ a random walk on the state space $\mathbb{Z}$ (Markov chain) with initial distribution $P(X_0)$ and transition probability $P(X_{t+1} = i + 1 | X_t = i) = p$, $P(X_{t+1} = i - 1 | X_t = i) = 1 - p$, and zero otherwise (here $0 < p < 1$). We suppose that we have ”observations” $Y_t$ of the state at time $t$ given by the output of an additive Gaussian noise channel:

$$Y_t = X_t + \sigma \xi_t, \quad t = 0, 1, 2$$

where $\xi_t \sim \mathcal{N}(0, 1)$ is Gaussian of mean zero and variance 1. The setting corresponds to the belief network of a Hidden Markov Model (Figure 1).

![Figure 1: Belief Network](image)

1. (4 pts) Write down the joint probability distribution of the whole belief network.

2. (4 pts) Are $Y_0$ and $Y_2$ independent random variables when conditioned on $X_1$? Are they independent when we do not condition? (no calculation but justification required).

3. (2 pts) Convert the belief network to a Markov Random Field and identify the maximal cliques, the corresponding factors, and the normalization factor $Z$.

4. (2 pts) From now on we initialize the Markov chain at time $t = 0$ with $X_0 = 0$. What is the initial distribution $P(X_0)$? And what is the effective alphabet (or possible values) of the random variables $X_1, X_2, Y_1, Y_2, Y_3$?

5. For this question the initialization is again $X_0 = 0$. We consider the Factor Graph representation of Figure 2.

   a) (6 pts) Set up the message passing equations and compute the marginal $\mu(Y_2)$ from those (see the recap of message passing equations below if needed). Express the result explicitly in terms of $p$ and $\sigma$.

   b) (2 pts) Do you think this calculation gives the exact marginal? Say why.

RECAP: Message passing equations for a general factor graph model $p(x) \propto \prod_a f_a(\{x_j : j \in \partial a\})$:

$$\mu_{i \rightarrow a}(x_i) = \prod_{b \in \partial a \setminus a} \mu_{b \rightarrow i}(x_i), \quad \mu_{a \rightarrow i}(x_i) = \sum_{x_j : j \in \partial a \setminus i} f_a(\{x_j : j \in \partial a\}) \prod_{j \in \partial a \setminus i} \mu_{j \rightarrow a}(x_j)$$
A leaf node is initialized with $\mu_{i\rightarrow a}(x_i) = 1$ and marginals are given by $\mu_i(x_i) \propto \prod_{a \in \partial i} \mu_{a\rightarrow i}(x_i)$. 
Problem 4. Tensor methods (20 pts)

Let \([\mu_1, \ldots, \mu_k]\) a set of \(k\) linearly independent column vectors of dimension \(n\) (with real components). We will assume throughout that these vectors have unit norm. Set

\[
T_2 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i, \quad T_3 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i
\]

where \(w_i, i = 1, \ldots, k, \) are real nonzero values.

We are given the arrays of components \(T_2^{\alpha \beta}, T_3^{\alpha \beta \gamma}, \alpha, \beta, \gamma \in \{1, \ldots, n\}\) and want to determine \(w_1, \ldots, w_k\) and \([\mu_1, \ldots, \mu_k]\). This problem guides you through a method that uses the simultaneous diagonalization of appropriate matrices.

The following multilinear transformation of \(T_3\) will be used

\[
T_3(I,I,u) = \sum_{i=1}^{k} w_i (\mu_i \otimes \mu_i)(u^T \mu_i)
\]

where \(I\) denotes the identity matrix and \(u\) an \(n\)-dimensional real column vector, \(u^T\) the transposed vector.

1. (7 pts) Let \(V = [\mu_1, \ldots, \mu_k]\) a square matrix. Show that

\[
T_2 = V \text{Diag}(w_1, \ldots, w_k)V^T, \quad T_3(I,I,u) = V \text{Diag}(w_1, \ldots, w_k) \text{Diag}(u^T \mu_1, \ldots, u^T \mu_k)V^T
\]

where \(\text{Diag}(a_1, \ldots, a_k)\) is the diagonal matrix with \(a_i\)'s on the diagonal.

2. (2 pts) Now we specialize to \(n = k\). Why is \(T_2\) an invertible matrix?

3. We choose \(u\) from a continuous distribution over \(\mathbb{R}^n\). Still in the case \(n = k\).

   a) (7 pts) Explain how to uniquely recover almost surely the set of \(\mu_i\)'s from the matrix

   \[
   M = T_3(I,I,u)T_2^{-1}
   \]

   using standard linear algebra methods.

   b) (4 pts) How do you then recover the \(w_i\)'s?
Problem 5. Multiple choice questions (20 pts)
Circle the correct answers. No justification required

1. (5 pts) [Several correct answers possible.] Let \( H = \{ h_\theta \}_{\theta \in \Theta} \) be a hypothesis class such that \( \text{VCdim}(H) = +\infty \). Then the set of parameters \( \Theta \):
   
   A. is finite.
   
   B. can be countable.
   
   C. can be uncountable.
   
   D. can be finite, countable or uncountable.

2. (5 pts) [Several correct answers possible.] Let \((x_i, y_i) \in \mathbb{R} \times \{0,1\}\) for \(i \in \{1, \ldots, n\}\). Let \( \hat{y}_i(w) = 1/(1 + e^{-wx_i}) \). Define
   
   \[
   f : w \in \mathbb{R} \mapsto -\sum_{i=1}^{n} \left[ y_i \log(\hat{y}_i(w)) + (1 - y_i) \log(1 - \hat{y}_i(w)) \right] + \lambda |w| ,
   \]
   
   where \( \lambda > 0 \). The function \( f \) is:
   
   A. convex.
   
   B. differentiable everywhere.
   
   C. subdifferentiable everywhere.
   
   D. Lipschitzian.

3. (5 pts) [Single correct answer.] According to the Hammersley-Clifford theorem the MRF property for a probability distribution \( p(x) > 0 \) implies
   
   \[
   p(x) = \frac{1}{Z} \prod_{\text{maximal cliques } C} \psi_C(\{x_i, i \in C\})
   \]
   
   where \( \psi_C(\{x_i, i \in C\}) > 0 \) and \( Z \) is the normalization factor. This decomposition is unique (up to the absorption of \( Z \) into factors):
   
   A. always.
   
   B. never.
   
   C. only when the MRF comes from a Belief Network.
   
   D. only if the graph of the MRF is a tree.
4. (5 pts) [Single correct answer.] Let \( w_i(\epsilon), i = 1, \cdots, K \) be continuous functions of \( \epsilon \in [0, 1] \). Let also \([a_1 + \epsilon a'_1, \cdots, a_K + \epsilon a'_K], [b_1 + \epsilon b'_1, \cdots, b_K + \epsilon b'_K], [c_1 + \epsilon c'_1, \cdots, c_K + \epsilon c'_K]\) be \( N \times K \) rank-\( K \) matrices for all \( \epsilon \). Consider the tensor

\[
T(\epsilon) = \sum_{i=1}^{K} w_i(\epsilon) (a_i + \epsilon a'_i) \otimes (b_i + \epsilon b'_i) \otimes (c_i + \epsilon c'_i)
\]

A. The tensor rank always equals \( K \) for all \( \epsilon \in [0, 1] \).

B. The tensor rank equals \( K \) for all \( \epsilon \in [0, 1] \) such that \( w_i(\epsilon) \neq 0, i = 1, \cdots, K \).

C. When we take a limit \( \lim_{\epsilon \to 0} T(\epsilon) \) it may happen that the tensor rank of the limit is \( K + 1 \).

D. If we replace the assumption that \([c_1 + \epsilon c'_1, \cdots, c_K + \epsilon c'_K]\) is rank \( K \), by the assumption that these vectors are pairwise independent, then the tensor rank can never be \( K \) whatever we assume for \( w_i(\epsilon), i = 1, \cdots, K \).